# Gradient-based optimization

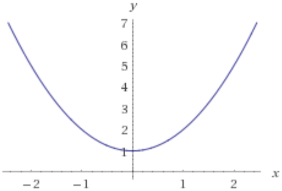
# Dérivative and extrema

**Theorem** If $f$ is differentiable in the vicinity of $a$, and has a local extremum in $a$, then $f'(a) = 0$.

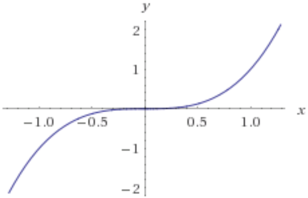
# Dérivative and extrema

**Theorem** If $f$ is differentiable in the vicinity of $a$, and has a local extremum in $a$, then $f'(a) = 0$.

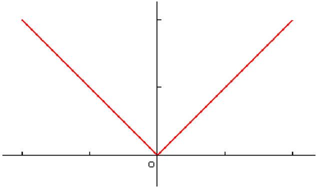
**AND NOTHING MORE !!**




(loc. extr. + diff.) $\Longrightarrow (f'(a) = 0)$




$(f'(a) = 0) \not\Longrightarrow$ loc. extr.




loc. extr. $\not\Longrightarrow (f'(a) = 0)$

# Necessary condition for an optimum

$$\text{Let} \qquad J: \begin{array}{rcl} E \subset \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ \mathbf{x} = (x_1, \ldots, x_n) & \longrightarrow & J(x_1, \ldots, x_n) \end{array}$$

- If $\widehat{x}$ is an internal point of $E$ (i.e. if there exists an open set $\Omega$ such that $\widehat{x} \in \Omega \subset E$) and if $J$ is differentiable in $\widehat{x}$, then

$$\widehat{x} \text{ local minimum of } J \quad \Longrightarrow \quad \nabla J(\widehat{x}) = 0$$

- If $E$ is convex, if $J$ is convex, then

$$\widehat{x} \text{ local minimum of } J \quad \Longleftrightarrow \quad \nabla J(\widehat{x}) = 0$$

# Necessary condition for an optimum

$$\text{Let} \quad J: \begin{array}{rcl} E \subset \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ \mathbf{x} = (x_1, \ldots, x_n) & \longrightarrow & J(x_1, \ldots, x_n) \end{array}$$

- If $\widehat{x}$ is an internal point of $E$ (i.e. if there exists an open set $\Omega$ such that $\widehat{x} \in \Omega \subset E$) and if $J$ is differentiable in $\widehat{x}$, then
$$\widehat{x} \text{ local minimum of } J \quad \Longrightarrow \quad \nabla J(\widehat{x}) = 0$$

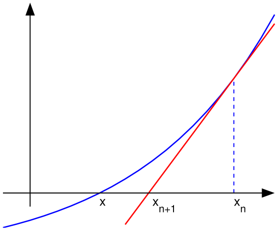
- If $E$ is convex, if $J$ is convex, then
$$\widehat{x} \text{ local minimum of } J \quad \Longleftrightarrow \quad \nabla J(\widehat{x}) = 0$$

$$\boxed{\text{Minimizing } J \quad \longrightarrow \quad \text{finding the roots of } \nabla J}$$

# Descent methods

Descent methods for minimizing the cost function require the knowledge of (an estimate of) its gradient.

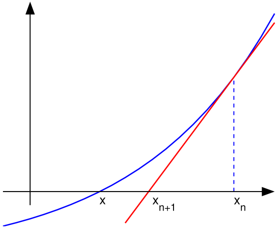
**1-D: Newton iteration to find the roots of a function $f$:** $x_{n+1} = ??$



# Descent methods

Descent methods for minimizing the cost function require the knowledge of (an estimate of) its gradient.

**1-D: Newton iteration to find the roots of a function $f$:** $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

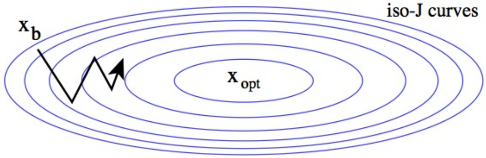



for the optimization: $f \equiv \nabla J$

# Descent methods

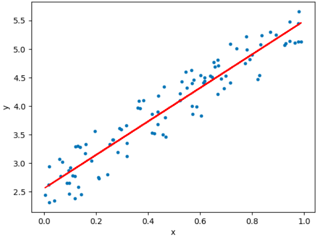
Descent methods for minimizing the cost function require the knowledge of (an estimate of) its gradient.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \, \mathbf{d}_k$$



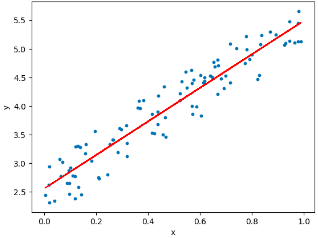

$$\text{with } \mathbf{d}_k = \begin{cases} -\nabla J(\mathbf{x}_k) & \text{gradient method} \\ -\left[\text{Hess}(J)(\mathbf{x}_k)\right]^{-1} \nabla J(\mathbf{x}_k) & \text{Newton method} \\ -\mathbf{B}_k \, \nabla J(\mathbf{x}_k) & \text{quasi-Newton methods (BFGS, \ldots)} \\ -\nabla J(\mathbf{x}_k) + \frac{\|\nabla J(\mathbf{x}_k)\|^2}{\|\nabla J(\mathbf{x}_{k-1})\|^2} d_{k-1} & \text{conjugate gradient} \\ \ldots & \ldots \end{cases}$$

# Least squares method: linear regression




Given some data $(x_i, y_i)_{i=1,\dots,p}$, what is the best approximate relationship $Y = aX + b$ ?

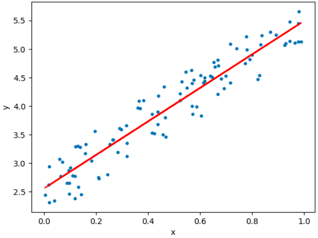
# Least squares method: linear regression




Given some data $(x_i, y_i)_{i=1,\ldots,p}$, what is the best approximate relationship $Y = aX + b$ ?

Let $y_i = ax_i + b + \varepsilon_i$ and minimize $E(a, b) = \sum_{i=1}^{p} \varepsilon_i^2 = \sum_{i=1}^{p} (y_i - ax_i - b)^2$

# Least squares method: linear regression




Given some data $(x_i, y_i)_{i=1,\ldots,p}$, what is the best approximate relationship $Y = aX + b$ ?

Let $y_i = ax_i + b + \varepsilon_i$ and minimize $E(a,b) = \sum_{i=1}^{p} \varepsilon_i^2 = \sum_{i=1}^{p} (y_i - ax_i - b)^2$

$$\begin{cases} \dfrac{\partial E}{\partial a}(\hat{a}, \hat{b}) = -2\displaystyle\sum_{i=1}^{p} x_i\,(y_i - \hat{a}\,x_i - \hat{b}) = 0 \\ \dfrac{\partial E}{\partial b}(\hat{a}, \hat{b}) = -\displaystyle\sum_{i=1}^{p} (y_i - \hat{a}\,x_i - \hat{b}) = 0 \end{cases}$$

i.e. $$\begin{cases} \left(\displaystyle\sum_{i=1}^{p} x_i\right)\hat{a} + p\,\hat{b} = \displaystyle\sum_{i=1}^{p} y_i \\ \left(\displaystyle\sum_{i=1}^{p} x_i^2\right)\hat{a} + \left(\displaystyle\sum_{i=1}^{p} x_i\right)\hat{b} = \displaystyle\sum_{i=1}^{p} x_i\,y_i \end{cases}$$

Hence $$\hat{a} = \frac{\dfrac{1}{p}\displaystyle\sum_{i=1}^{p} x_i y_i - \bar{x}\bar{y}}{\dfrac{1}{p}\displaystyle\sum_{i=1}^{p} (x_i - \bar{x})^2}$$ and $$\hat{b} = \bar{y} - \hat{a}\,\bar{x}$$

# Generalization: minimum of a quadratic function in finite dimension

**Theorem: Generalized (or Moore-Penrose) inverse**

Let $\mathbf{M}$ a $p \times n$ matrix, with rank $n$, and $\mathbf{b} \in \mathbb{R}^p$. *(hence $p \geq n$)*

Let $J(\mathbf{x}) = \|\mathbf{M}\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{M}\mathbf{x} - \mathbf{b})^T(\mathbf{M}\mathbf{x} - \mathbf{b})$.

$J$ is minimum for $\hat{\mathbf{x}} = \mathbf{M}^+\mathbf{b}$ , where $\mathbf{M}^+ = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ (*generalized inverse*, or *Moore-Penrose inverse*).

Exercise : prove it.

# Generalization: minimum of a quadratic function in finite dimension

**Theorem: Generalized (or Moore-Penrose) inverse**

Let $\mathbf{M}$ a $p \times n$ matrix, with rank $n$, and $\mathbf{b} \in \mathbb{R}^p$. *(hence $p \geq n$)*

Let $J(\mathbf{x}) = \|\mathbf{M}\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{M}\mathbf{x} - \mathbf{b})^T(\mathbf{M}\mathbf{x} - \mathbf{b})$.

$J$ is minimum for $\hat{\mathbf{x}} = \mathbf{M}^+\mathbf{b}$ , where $\mathbf{M}^+ = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ (*generalized inverse*, or *Moore-Penrose inverse*).

Exercise : check that this is consistent with the previous results of linear regression.

# Generalization: minimum of a quadratic function in finite dimension

**Theorem: Generalized (or Moore-Penrose) inverse**

Let $\mathbf{M}$ a $p \times n$ matrix, with rank $n$, and $\mathbf{b} \in \mathbb{R}^p$. *(hence $p \geq n$)*

Let $J(\mathbf{x}) = \|\mathbf{Mx} - \mathbf{b}\|^2 = (\mathbf{Mx} - \mathbf{b})^T(\mathbf{Mx} - \mathbf{b})$.

$J$ is minimum for $\boxed{\hat{\mathbf{x}} = \mathbf{M}^+\mathbf{b}}$, where $\mathbf{M}^+ = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ (*generalized inverse*, or *Moore-Penrose inverse*).

**Corollary: with a generalized norm**

Let $\mathbf{N}$ a $p \times p$ symmetric definite positive matrix.

Let $J_1(\mathbf{x}) = \|\mathbf{Mx} - \mathbf{b}\|_N^2 = (\mathbf{Mx} - \mathbf{b})^T\mathbf{N}\,(\mathbf{Mx} - \mathbf{b})$.

$J_1$ is minimum for $\boxed{\hat{\mathbf{x}} = (\mathbf{M}^T\mathbf{N}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{N}\,\mathbf{b}}$.

# Just for scientific culture: optimization in infinite dimension, using the adjoint

$$\begin{cases} -u''(x) + c(x)\,u'(x) = f(x) & x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 \end{cases}$$

- $c(x)$ is unknown
- $u^{\text{obs}}(x)$ is an observation of $u(x)$

**Problem** How can we estimate the function $c(x)$?

# Just for scientific culture: optimization in infinite dimension, using the adjoint

$$\begin{cases} -u''(x) + c(x)\,u'(x) = f(x) & x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 \end{cases}$$

- $c(x)$ is unknown
- $u^{\text{obs}}(x)$ is an observation of $u(x)$

**Problem** How can we estimate the function $c(x)$?

**Idea** minimize the cost function: $J(c) = \dfrac{1}{2}\displaystyle\int_0^1 \left(u(x) - u^{\text{obs}}(x)\right)^2 \, dx$

$\longrightarrow$ need to compute $\nabla J(c)$

# Example of the adjoint method

$$\left\{\begin{array}{ll} -u''(x) + c(x)\,u'(x) = f(x) & \qquad x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 & \end{array}\right.$$

Find $c(x)$ that minimizes $\quad J(c) = \dfrac{1}{2}\displaystyle\int_0^1 \left(u(x) - u^{\mathrm{obs}}(x)\right)^2 dx$

# Example of the adjoint method

$$\begin{cases} -u''(x) + c(x)\,u'(x) = f(x) & x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 \end{cases}$$

$$\text{Find } c(x) \text{ that minimizes} \quad J(c) = \frac{1}{2}\int_0^1 \left(u(x) - u^{\text{obs}}(x)\right)^2 dx$$

$\nabla J \to$ Gâteaux-derivative: $\hat{\mathrm{J}}[c](\delta c) = < \nabla J(c), \delta c >$

$$\hat{\mathrm{J}}[c](\delta c) = \int_0^1 \hat{u}(x)\left(u(x) - u^{\text{obs}}(x)\right) dx \quad \text{with } \hat{u} = \lim_{\alpha\to 0} \frac{u_{c+\alpha\delta c} - u_c}{\alpha}$$

What is the equation satisfied by $\hat{u}$ ?

$$\begin{cases} -\hat{u}''(x) + c(x)\,\hat{u}'(x) = -\delta c(x)\,u'(x) & x \in ]0,1[ \\ \hat{u}(0) = \hat{u}(1) = 0 \end{cases}$$

tangent
linear model

# Example of the adjoint method

$$\begin{cases} -u''(x) + c(x)\,u'(x) = f(x) & x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 \end{cases}$$

$$\text{Find } c(x) \text{ that minimizes} \quad J(c) = \frac{1}{2}\int_0^1 \Big(u(x) - u^{\text{obs}}(x)\Big)^2\, dx$$

$\nabla J \rightarrow$ Gâteaux-derivative: $\hat{J}[c](\delta c) = < \nabla J(c), \delta c >$

$$\hat{J}[c](\delta c) = \int_0^1 \hat{u}(x)\,\Big(u(x) - u^{\text{obs}}(x)\Big)\, dx \quad \text{with } \hat{u} = \lim_{\alpha\to 0} \frac{u_{c+\alpha\delta c} - u_c}{\alpha}$$

What is the equation satisfied by $\hat{u}$ ?

$$\begin{cases} -\hat{u}''(x) + c(x)\,\hat{u}'(x) = -\delta c(x)\,u'(x) & x \in ]0,1[ \\ \hat{u}(0) = \hat{u}(1) = 0 \end{cases}$$

tangent
linear model

Going back to $\hat{J}$ scalar product of the TLM with a variable $p$: $-\int_0^1 \hat{u}''p + \int_0^1 c\,\hat{u}'p = -\int_0^1 \delta c\, u'p$

Integration by parts: $\int_0^1 \hat{u}\,(-p'' - (c\,p)') = \hat{u}'(1)p(1) - \hat{u}'(0)p(0) - \int_0^1 \delta c\, u'p$

# Example of the adjoint method

$$\begin{cases} -u''(x) + c(x)\,u'(x) = f(x) & x \in ]0,1[,\ f \in L^2(]0,1[) \\ u(0) = u(1) = 0 \end{cases}$$

Find $c(x)$ that minimizes $\quad J(c) = \dfrac{1}{2}\displaystyle\int_0^1 \left(u(x) - u^{\text{obs}}(x)\right)^2 dx$

$\nabla J \rightarrow$ Gâteaux-derivative: $\hat{\mathrm{J}}[c](\delta c) = < \nabla J(c), \delta c >$

$$\hat{\mathrm{J}}[c](\delta c) = \int_0^1 \hat{u}(x)\left(u(x) - u^{\text{obs}}(x)\right) dx \quad \text{with } \hat{u} = \lim_{\alpha \to 0} \frac{u_{c+\alpha\delta c} - u_c}{\alpha}$$

What is the equation satisfied by $\hat{u}$ ?

$$\begin{cases} -\hat{u}''(x) + c(x)\,\hat{u}'(x) = -\delta c(x)\,u'(x) & x \in ]0,1[ \\ \hat{u}(0) = \hat{u}(1) = 0 \end{cases}$$

tangent linear model

Going back to $\hat{\mathrm{J}}$ scalar product of the TLM with a variable $p$: $-\displaystyle\int_0^1 \hat{u}''p + \int_0^1 c\,\hat{u}'p = -\int_0^1 \delta c\,u'p$

Integration by parts: $\displaystyle\int_0^1 \hat{u}\left(-p'' - (c\,p)'\right) = \hat{u}'(1)p(1) - \hat{u}'(0)p(0) - \int_0^1 \delta c\,u'p$

$$\begin{cases} -p''(x) - (c(x)\,p(x))' = u(x) - u^{\text{obs}}(x) & x \in ]0,1[ \\ p(0) = p(1) = 0 \end{cases}$$

adjoint model

Conclusion: $\nabla J(c(x)) = -u'(x)\,p(x)$